National Endowment for the Humanities - Office of Digital Humanities

White Paper

Grant Number: HD 248600-16

Project Title: "Reading Chicago Reading: Modeling Texts and Readers in a Public Library System"

Director: John Shanahan

Institution: DePaul University

Submitted: March 31, 2019

Project Activities and Accomplishments

During the project period of June 2016 to December 2018, members of the RCR project team published three peer-reviewed papers (and more are in the pipeline); presented about the project over a dozen times at professional meetings across the U.S. and around the world; trained students in useful work skills by way of project goals; and ultimately created a first version of the predictive model that we sought to create with the support of the NEH. This work happened in collaboration with students and colleagues at DePaul and the Chicago Public Library.

We accomplished the major goal of the research plan, but during the two and a half years funded by the grant we also developed several additional directions for research and new partners and interests (the details are included below). The support of the NEH ODH was a true catalyst that launched a major DH project from its opening work to maturity of results. The influence of the project will become manifest as our published work expands the influence of our work.

In the original project plan submitted to the NEH ODH in 2015, we proposed a hypothesis that "text characteristics, library branch demographics, and promotional activities are variables that can be used to predict patron response to an OBOC title. Our primary task will be to discover the relationships between these variables and encode them in predictive models." This work has largely been done, and late in 2018 we tested the predictive model for seven recent seasons of the CPL OBOC program. We hope to test it soon with the short list and chosen title of the next (i.e. 2019/20) OBOC season before the announcement of the book title to the public in fall 2019. Doing this enables us not only to establish the accuracy of the model but also to provide a data-driven insight into the reading life of the city of Chicago to the Chicago Public staff and larger public.

Two changes of key personnel took place during the grant period: in fall 2016 we added Dr. Ana Lucic, DePaul's Digital Scholarship Librarian and an expert in text mining, to the project. In early 2018, co-PI Megan Bernal, Associate University Librarian, left DePaul university for the Los Alamos National Laboratory Library. Because Dr. Lucic is familiar with library systems and the HathiTrust (two key elements for our work), the departure of Ms. Bernal near the close of the grant period did not adversely affect our work.

During the bulk of the grant period, the project co-PIs met weekly to discuss work in progress and determine new goals. At these meetings, students and associated colleagues also contributed ideas and work in progress. Over the course of the grant, nine graduate students and three undergraduates worked on project tasks ranging from archiving and interviewing to hand correction of data, visualizations, and code development. Our project work featured outreach to additional partners that we had not anticipated originally. For example, in 2017, we were fortunate to win a HathiTrust Research Center Advanced Computing Support Grant that allowed us to work on the in-copyright texts which represent the majority of our OBOC corpus. The work

conducted through the HathiTrust secure data capsule has become part of already published work (see JCDL 2019) and will be featured in at least two forthcoming articles.

More recently, in 2018, we received a Lyrasis Catalyst grant to develop a dashboard to visualize the findings of our research. As a result, our meetings have become a productive environment for graduate students to work on the project in consultation with organizations beyond campus -- this grant furthers education and professionalization of the students while also expanding project goals.

Our project has benefited greatly from the involvement of Nandhini Gulasingam, a specialist in GIS mapping at DePaul. Ms. Gulasingam helped us make a series of maps several times during the grant period, and the results of her work -- particularly in connection with Burke, Lucic, and Stoica -- are included below. A number of presentations and forthcoming papers are indebted to her collaboration. While extensive mapping was not fully described in the original proposal to the NEH ODH, over the grant period we identified a number of ways to use GIS to enhance the explanatory power and meaning of our data (for example, tying patterns of sentiment in our text corpus to Gini coefficient measures of Chicago's urban inequality was not a feature of our early planning, but it has shown promise).

Over the grant period project team members delivered over a dozen major presentations and many smaller ones in venues as varied as the English department at DePaul, senior leadership of the Chicago Public Library, and national leadership meetings of Lyrasis and the HathiTrust. The list of major presentations is below. Our website (https://dh.depaul.press/reading-chicago/) and Twitter account (@readchireading) have been useful ways to publicize results and make findings and code available. We have included a few pictures and screenshots documenting this outreach.

Audiences

Our project serves several audiences. Two key groups are scholars in the digital humanities and public librarians. Subsidiary audiences are book groups, fan-fiction sites, and publishers – and websites dedicated to reading and writing such as Goodreads and Archive of Our Own.

Our research is of interest, and use, to DH scholars for a number of reasons. First is the novelty of our approach, mixing five major kinds of data (some elements of which repeatedly 'refresh,' as it were) in order to create a predictive model about a city-size cultural program. Another benefit to DH scholars is open-source software tools that may be of use in their own context: these include a new means of detecting textual boundaries (see JCDL 2019 paper) and -- still in progress -- new means to quantify sentiment scores associated with locations in texts. In the case of the latter, we expect that our findings will be in dialogue with the fascinating work of Archer and Jockers in their *Bestseller Code* (2016).

The results of our project will be of particular interest to library consortia and librarians in general because the results point to the factors that play a role when planning a special promotion of a work or an author. The predictive model that we prototyped at the end of 2018 on the known set of recent OBOC books can now be run on new, non-OBOC books. Our goal of city-wide branch circulation predictions is now in its beta stage, and we expect in the coming months (spring/summer 2019), before the book title is made public, to predict its circulation across the city. The model that we have built and the results of the analyses that we have run inform the libraries -- and in particular a consortium or a system of public libraries in a major American city -- of the effects of a concentrated and wide-scale promotion and their impact on the circulation patterns. Probably most importantly, our model can be replicated if the main predictor variables are available: in particular, library circulation statistics prior to an event, book features, as well as a record of promotional activities. With these features in place, a similar model can be built and librarians can obtain data-driven insights about the choices they make.

Although we do not have concrete metrics of direct impact, our presentations have had an estimated audience of several hundred people in total in the U.S., Canada, Taiwan, and Japan. Project publications will reach an even wider audience. Publicity about our project in academic circles has given some additional currency to the work of the Chicago Public Library and helped deepen DePaul's reputation for work in digital humanities (where it is still somewhat nascent).

Evaluation

There have been a number of evaluations of the work as peer-reviewed articles and presentations. We have, in addition, received grant support to continue and to push into new directions. Admittedly, the project as a whole has not yet been evaluated -- but the approval and enthusiasm of Chicago public librarians with respect to our findings are encouraging.

A major strength of the project is the insight into integrating otherwise separate streams of data and using them in a single model that can give quantitative insight about resource allocation to generally underfunded public library systems. We have modeled, and described in our presentations and papers, ideas as well as outlined the challenges of conducting this kind of work.

A weakness of our project is the small corpus of CPL circulation records (totaling at first just seven books but increasing to over 76 by the end) most of which are in-copyright works. This meant that book features were hard to obtain (due to copyright laws) and that our analyses were restricted to non-consumptive type of analysis. We had extra steps in text processing such as supplying three books to the HathiTrust Data Capsule ourselves in order to be able to have each of the seven OBOC selections since 2011 in the digital library. To scale this project outward to greater predictive insight will require a larger set of books amenable to text mining as well as a fuller access to library circulation records.

Although limited, coverage of the project in the public sphere has been positive, for example a write up of the project in *Library Journal* in 2016 and exposure of the project to a nation-wide library administration audience at the Lyrasis Leadership Summit in late 2018. At academic conferences we have received expressions of interest and constructive feedback.

Continuation and Long-Term Impact of the Project

Although the grant has expired, we have been able to secure additional sources of funding and maintain the intellectual momentum of the work. Our research group continues to meet regularly. Our key deliverable to be made developed with the NEH ODH's support, a predictive circulation model, can now be used to inform the work of library staff at the Chicago Public Library. The NEH ODH support allowed us the research time and financial support to build the architecture for the research and its propagation.

We have collected a great amount of data in the form of extracted features from the works (e.g. locations, sentiments, readability measures), Twitter posts about One Book One Chicago events, circulation statistics as provided by Chicago Public Library, as well as Goodreads reviews for the OBOC selections. Although some preliminary analysis of these datasets has already been performed, quite a few new research avenues have opened up that we plan to continue working on in the future. Additionally, during the process of problem-solving for the central task of modeling, a number of related topics and methods and new research areas have emerged that we think will be of interest to other researchers: a method for determining main text in a work and separating it from paratextual boundaries within the context of the HathiTrust digital library; novel ways of mapping literary works according to geographical terms and sentiment scores; means of aligning public library programming, social media, and book checkout figures; as well as a few other research threads.

Award Products

Peer-Reviewed Papers

1. "Circulation Modeling of Library Book Promotions." Co-authors: Robin Burke, Ana Lucic, and John Shanahan. *Proceedings of the ACM Joint Conference on Digital Libraries*. Toronto, Canada (June 2017)
2. "Real and Imagined Geography at City-Scale: Sentiment Analysis of Chicago's 'One Book' Program." Co-authors: Ana Lucic and John Shanahan. DH2017 Annual Conference. Montreal, Canada (August 2017)
3. "Unsupervised Clustering with Smoothing for Detecting Paratext Boundaries in Scanned Documents." Co-authors: Ana Lucic, Robin Burke, and John Shanahan. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. Urbana-Champaign, IL. (June 2019)

Peer-Reviewed Papers Submitted

1. “Reading Chicago Reading: Quantitative Analysis of a Repeating Literary Program.” Co-authors: John Shanahan, Robin Burke, and Ana Lucic. Under review at *Digital Humanities Quarterly*.

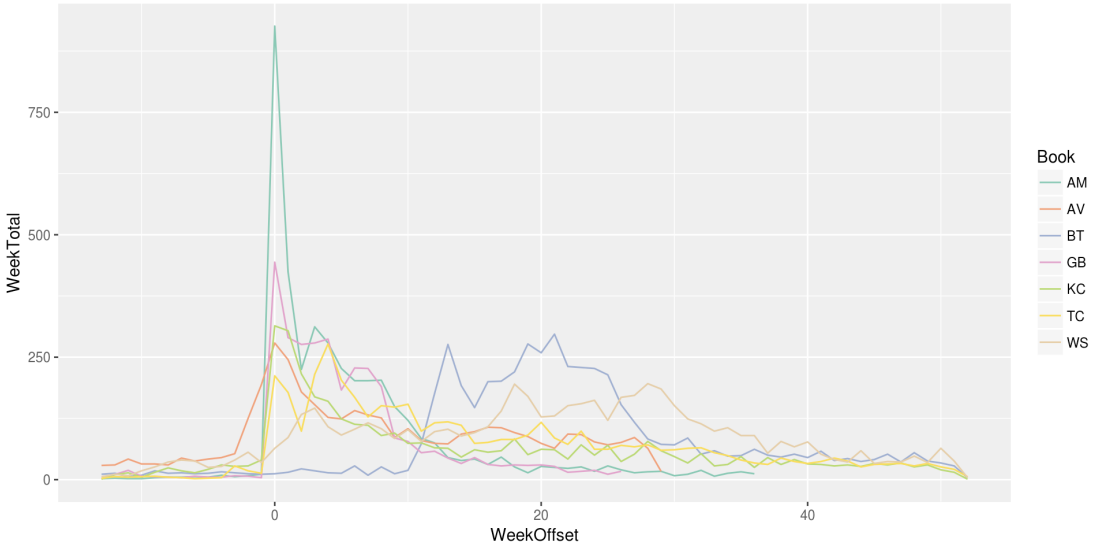
Workshop
1. “Mining Diverse Texts for Location and Sentiment.” Robin Burke, Ana Lucic, and John Shanahan. Chicago Colloquium for Digital Humanities and Computer Science. University of Illinois at Chicago (November 2017)

Major Presentations
1. “Data Sources for Modeling Library Texts and Readership.” Burke, Shanahan, and Budde. 2nd Annual International Conference on Computational Social Science. Northwestern University, Evanston, IL. (June 2016)
2. “Visualizing and Modeling Chicago’s ‘One Book’ Program.” LITA: Library and Information Technology Association Forum. Burke, Shanahan, and Bernal. Fort Worth, TX. (November 2016)
3. “Reading Chicago Reading: Capture and Analysis of City-scale Literary Events.” Shanahan, Burke, Lucic. Chicago Colloquium on Digital Humanities and Computer Science, University of Illinois of Chicago. (November 2016)
4. “Exploratory Search Beyond the Work Level.” Luedtke, Lucic, and Bernal. ALCTS Exchange. Online (May 2017).
5. "Real and Imagined Geography at City-Scale: Sentiment Analysis of Chicago’s ‘One Book’ Program" Lucic, Shanahan. Digital Humanities 2017. Montreal, Canada. (August 2017).
6. “DH Tools to Capture and Predict Literary Reading at City-Scale.” Shanahan. Eighth International Conference on Digital Archives and Digital Humanities. Taipei, Taiwan. (December 2017).
7. “The Geography of Circulation and Sentiment: Mapping ‘One Book One Chicago.’” Lucic, Gulasingam, Shanahan. Chicago Colloquium on Digital Humanities and Computer Science. Chicago. (November 2018)
8. “The ‘Reading Chicago Reading’ Project: Successes and Challenges Using HathiTrust for In-Copyright Corpora.” Shanahan, Lucic, Burke. HathiTrust Annual Meeting, Rosemont, IL. (October 2018).
9. “Capturing Literary Events at Metropolitan Scale: Open Data and ‘One Book One Chicago.’” Shanahan. Japanese Association for Digital Humanities 8th Annual Conference, Tokyo, Japan. (September 2018).
10. “Mapping City-Scale Reading Events: Geography and Sentiment of ‘One Book One Chicago.’” Lucic, Shanahan, Stoica. Association for Computers and the Humanities. Pittsburgh. (July 2019).

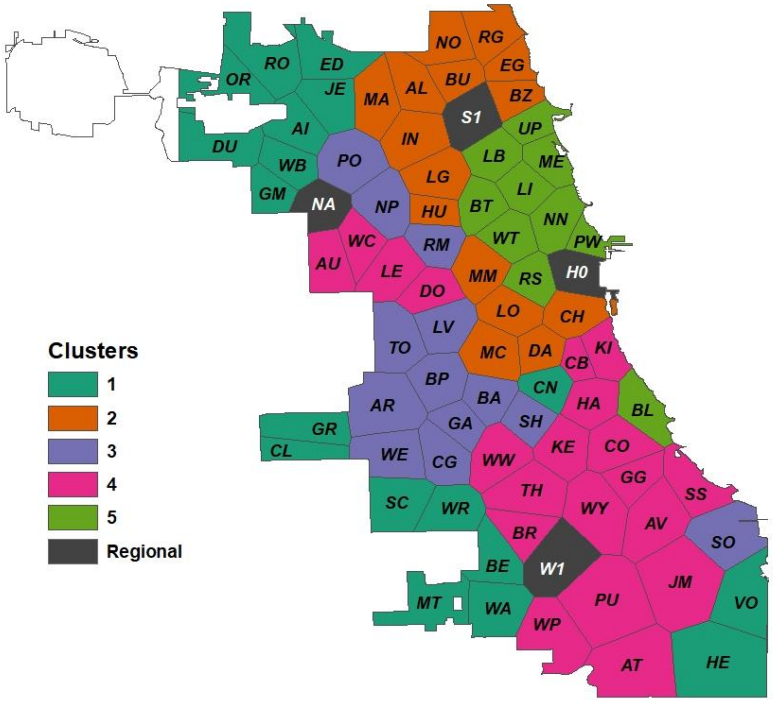
Other Grants Received (June 2016-December 2018)

1. Microsoft Azure (2015-16)
2. HTRC ACS (2017)
3. Lyrasis (2018-19)

Selected data and figures from project papers published and forthcoming

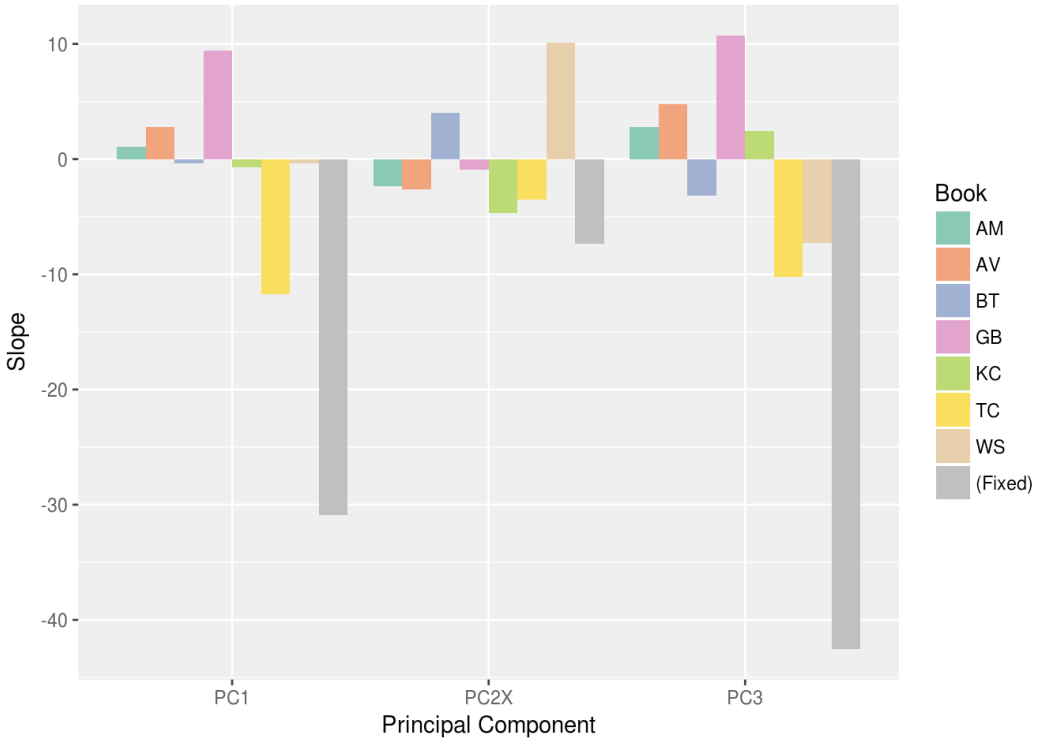


Book circulation totals (all branches) for seven seasons of OBOC (2011-2017) superimposed with the official launch date set to zero on the x axis.



Chicago Public Library branches colored by cluster

Normalized branch circulation by book, colored by cluster.



Fitted coefficients demographic variables in the multi-level model

OBOC text measures

Title abbreviation	Number of words (punctuation excluded)	Average sentence length (punctuation excluded)	Dale-Chall Index	Type-token ratio	Combined difficulty
AM	263,427	16.17	8.03	8.61	0.43
GB	71,138	21.40	8.27	9.90	0.38
BT	127,838	10.58	8.08	7.34	0.08
WS	212,613	19.02	8.53	9.60	0.53
KC	240,216	16.89	8.89	9.20	0.54
TC	150,166	24.10	10.47	11.50	0.85
AV	125,849	18.67	9.50	11.11	0.60